Biometrics 62, 36–48
March 2006

DOI: 10.1111/j.1541-0420.2005.00442.x

Case–Control and Case-Only Designs with Genotype and Family History Data: Estimating Relative Risk, Residual Familial Aggregation, and Cumulative Risk

Nilanjan Chatterjee,[1,*] Zeynep Kalaylioglu,[2] Joanna H. Shih,[3] and Mitchell H. Gail[1]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, 6210 Executive Boulevard, Rockville, Maryland 20852, U.S.A.
[2]Information Management Services, 6110 Executive Boulevard, Rockville, Maryland 20852, U.S.A.
[3]Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, 6210 Executive Boulevard, Rockville, Maryland 20852, U.S.A.
[*]*email:* chattern@mail.nih.gov

Summary. In case–control studies of inherited diseases, participating subjects (probands) are often interviewed to collect detailed data about disease history and age-at-onset information in their family members. Genotype data are typically collected from the probands, but not from their relatives. In this article, we introduce an approach that combines case–control analysis of data on the probands with kin–cohort analysis of disease history data on relatives. Assuming a marginally specified multivariate survival model for joint risk of disease among family members, we describe methods for estimating relative risk, cumulative risk, and residual familial aggregation. We also describe a variation of the methodology that can be used for kin–cohort analysis of the family history data from a sample of genotyped cases only. We perform simulation studies to assess performance of the proposed methodologies with correct and misspecified models for familial aggregation. We illustrate the proposed methodologies by estimating the risk of breast cancer from BRCA1/2 mutations using data from the Washington Ashkenazi Study.

Key words: Ascertainment correction; Copula model; Kin–cohort; Multivariate survival; Penetrance.

1. Introduction

The case–control sampling design, which has been widely used in classical questionnaire-based epidemiologic studies, is now being increasingly used to examine genetic association and gene–environment interaction. This design, by collecting exposure information retrospectively on a fixed number of diseased (cases) and nondiseased subjects (controls), is efficient for studying the etiology of rare diseases. In case–control studies of genetic factors, participating subjects (probands) are often interviewed to collect detailed data about disease history and age-at-onset information in their family members. Genotype information is typically not available for the relatives of the cases and controls. Standard analysis of such case–control data involves fitting a logistic regression model to the case–control outcome with genotype and family history treated as covariates. From the classic results of Cornfield (1951), Andersen (1970), and Prentice and Pyke (1979), it is well known that this logistic regression analysis efficiently estimates the prospective odds ratios associated with the covariates, which approximate relative risks for rare diseases. The absolute risk of the disease cannot be estimated from this analysis without external information on the marginal probability of the disease in the population, which is needed to correct the logistic regression intercept. Moreover, treating family history as a covariate can create bias in parameter estimates due to inadequate adjustment for family size (Khoury and Flander, 1995).

For case–control studies where both disease history and covariate status are available on relatives, Whittemore (1995) and Zhao et al. (1998) developed likelihood- and estimating-equation-based methodology, respectively, both in the binary outcome setting, which uses all information from case–control probands and their relatives to produce efficient estimates of disease–covariate association and magnitude of familial aggregation. For a similar design, Li, Yang, and Schwartz (1998), Hsu et al. (1999), and Shih and Chatterjee (2002) developed different likelihood and pseudo-likelihood methodologies in the survival analysis setting that can account for censoring and age-at-onset information of disease. None of these methodologies, however, are directly applicable for the case–control study design we consider in the present article, where information on the covariate of interest, namely the genetic variant under study, is not available on the relatives.

Wacholder et al. (1998) developed a method to estimate gene-specific cumulative risk from analysis of "kin–cohort" data and the ages at disease onset, if any, among the probands' relatives. In this approach, the disease status and age at onset of the relatives are treated as outcome variables of interest in a survival analysis. The authors showed that although the relatives are not genotyped, one can estimate the

age-specific cumulative risk (penetrance) of the disease associated with different genotypes by linking the outcome data from the relatives to the genotype of the probands. Chatterjee and Wacholder (2001) developed a "marginal-likelihood" approach for analyzing kin–cohort data, which overcomes various limitations of the original analytic method of Wacholder et al. (1998), including the possibility of obtaining nonmonotone estimates of age-specific cumulative risk function.

Gail et al. (1999a, 1999b) called the case–control design with detailed family history information that collects genotype data on probands the "genotyped proband" design. They assumed that cases and controls were sampled randomly from all cases and controls in the population, respectively. The authors derived the likelihood of the genotypes of the probands and disease history data of the relatives conditional on the case–control status of the probands. They showed that this likelihood can be factored as the product of a traditional "case–control likelihood" of the genotype of probands given their case–control status and a "kin–cohort likelihood" for the relatives' disease outcome data given the genotype of the proband. Moore et al. (2001) used this likelihood framework to develop methods of parameter estimation for survival models. These authors noted that full maximum likelihood estimation using the true likelihood of the data can be computationally challenging and presented alternative pseudo-likelihood methods. These pseudo-likelihood approaches, however, fail to extract valuable relative risk information from case–control data and hence are inefficient.

Gail et al. (1999a, 1999b) and Chatterjee and Wacholder (2001) pointed out that the likelihood used by Gail et al. (1999a) is based on the assumption that all familial outcomes were conditionally independent given individuals' genotypes and that violation of this assumption of no "residual familial aggregation" can lead to biased parameter estimates. The effect of "residual familial aggregation" on cumulative risk estimation using disease incidence data of relatives of a case-enriched or even a case-only sample of subjects has been a matter of considerable debate in recent years (Begg, 2002; Gong and Whittemore, 2003).

The goal of the current article is to develop a combined approach of kin–cohort and case–control analysis that has the following strengths: (i) gives a computationally feasible method for extracting maximal information on relative risk and cumulative risk parameters from the data; (ii) relaxes the key assumption of "no residual familial aggregation" and quantifies the magnitude of such correlation; and (iii) automatically accounts for potential ascertainment bias in absolute risk estimation. In addition, these methodologies, with some modification, can be used for kin–cohort analysis of the family history data from a sample of genotyped cases only. We develop the methodology by extending and combining a number of estimation techniques that we have developed over a number of years in three different but related areas, namely kin–cohort analysis (Wacholder et al., 1998; Chatterjee and Wacholder, 2001), genotype-proband design (Gail et al., 1999a; Moore et al., 2001), and case–control family data (Shih and Chatterjee, 2002).

2. Methods

2.1 *Data Structure*

Consider a case–control study design where N_0 cases and N_1 controls have been randomly sampled from the cases and the controls in an underlying population, respectively. Let $Y_i^P = (T_i^P, \Delta_i^P)$, $i = 1, \ldots, N_0 + N_1$ denote the phenotype for $N_0 + N_1$ case–control subjects (probands). Here, if the ith subject is a case, $\Delta_i^P = 1$ and T_i^P denotes the age at onset of the disease and if the ith subject is a control $\Delta_i^P = 0$ and T_i^P denotes the age of the subject at interview. Suppose the ith proband reports disease history for M_i relatives and let $Y_{ij}^R = (T_{ij}^R, \Delta_{ij}^R)$, $j = 1, \ldots, M_i$ denote the phenotype history of these relatives. Here, if the jth relative has been reported to have the disease then $\Delta_{ij}^R = 1$ and T_{ij}^R denotes the corresponding reported age at onset and if the jth relative has been reported to be disease free then $\Delta_{ij}^R = 0$ and T_{ij} denotes the age of the relative at the time of interview of the proband or age at death if the relative had died before the study took place. Let G_i^P denote the genotype for the ith proband. Similarly, let G_{ij}^R denote the unobserved genotype for jth relative of the ith proband. Let $\boldsymbol{Y}_i^R = (Y_{i1}^R, \ldots, Y_{iM_i}^R)$ and $\boldsymbol{G}_i^R = (G_{i1}^R, \ldots, G_{iM_i}^R)$ denote the vector of phenotypes and genotypes, respectively, for the M_i relatives of proband i. For simplicity, we assume G is binary, indicating presence ($G = 1$) or absence ($G = 0$) of a dominant mutation. The proposed methodology, however, can be extended in a straightforward way for other types of genotype data. Let f denote the allele frequency for the mutation and $\pi = f^2 + 2f(1 - f)$ denote the carrier frequency, namely the probability of carrying the mutation.

2.2 *Model*

We assume our goal is to estimate risk parameters in marginal models for randomly selected gene carriers ($G = 1$) and noncarriers ($G = 0$). Let $\lambda_0(t)$ denote the age-specific hazard function of the disease among noncarriers. We allow $\lambda_0(t)$ to be completely unspecified (nonparametric). We will further assume the hazard function for carriers ($G = 1$) is given by a piecewise proportional hazard (PH) form

$$\lambda_1(t) = \lambda_0(t) \exp\{\beta(t)\} \quad \text{with}$$
$$\beta(t) = \sum_{l=0}^{K} \beta_l I(u_l \le t < u_{l+1}), \tag{1}$$

where $u_0 < u_1 < u_2 \cdots < u_{K+1}$ are a set of prespecified knots in the appropriate age range of interest. Although the traditional PH model assumes the proportionality factor β to be constant over all ages, we prefer the piecewise PH model that allows one to examine whether the effect of the gene varies by age. Hereafter, we will denote $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)^T$. Of course, the most flexible approach would be to assume $\lambda_1(t)$ to be completely unspecified, an approach that we have previously considered for kin–cohort estimation that ignores ascertainment (Chatterjee and Wacholder, 2001). Accounting for ascertainment, in the nonparametric hazard models, however, can be computationally challenging and may face an identifiability problem. In many survival analysis problems, where estimation of relative risk parameters are the focus of scientific interest, the baseline hazard function $\lambda_0(t)$ is treated as

a nuisance parameter, the estimation of which is avoided by constructing an appropriate partial likelihood that only depends on the relative risk parameters. In the context of this article, however, estimation of $\lambda_0(t)$ is essential for obtaining estimates of the cumulative risk parameters of interest.

To allow for residual familial aggregation given genotypes, we consider a model for specifying joint risks of the disease among the proband and his/her family members. We use copula models (Genest and McKay, 1986) which parameterize the joint risk of the disease in terms of the marginal risk for individual family members and additional dependence parameter(s) that characterize the correlations between the family members. Let $S_G(t) = \exp\{-\Lambda_G(t)\} = \exp\{-\int_0^t \lambda_G(s)\,ds\}$ denote the marginal survivor function for an individual relative in the family given his/her individual genotype (G), the hazard function $\lambda_G(s)$ being specified by the hazard model given in formula (1). In the copula model, the joint survivor function for a proband ($m = 0$) and his/her M family members ($m = 1, \ldots, M$) is specified as

$$\begin{aligned}&\Pr(T_0 > t_0, T_1 > t_1, \ldots, T_M > t_M \,|\, G_0, G_1, \ldots, G_M)\\ &\quad = C_\theta\{S_{G_0}(t_0), S_{G_1}(t_1), \ldots, S_{G_M}(t_M)\},\end{aligned} \tag{2}$$

where $C_\theta(u_1, \ldots, u_m)$, $\theta \in \Theta$ is a class of multivariate distribution functions defined on the product space of $[0, 1]^{M+1}$ with uniform marginal distributions. In this model, the parameter θ can be interpreted as a measure of "residual familial aggregation" that characterizes familial correlation of the disease that cannot be explained by the gene under study. The exact interpretation of θ, however, depends on the choice of the copula function $C_\theta(u_0, u_1, \ldots, u_M)$. For most of this article, we use Clayton's model (1978) that corresponds to the copula function

$$C_\theta(u_0, u_1, \ldots, u_m) = \left[\sum_{m=0}^{M} {u_m}^{1-\theta} - M + 1\right]^{1/(1-\theta)}. \tag{3}$$

Model (3) corresponds to constant value (θ) for the cross-ratio function, a measure of local dependence between pairs of survival times that was introduced by Oakes (1989). The value of $\theta = 1$ corresponds to independence and $\theta > 1$ corresponds to positive dependence. Although in a restricted range, values $\theta \leq 1$ can be allowed to model negative correlation, in this article we only allow for positive dependence ($\theta \geq 1$) for modeling familial aggregation.

Inference in copula models for randomly sampled multivariate units with censored survival times was described by Genest and Rivest (1993), Shih and Louis (1995), and Shih (1998). Estimation of parameters in our framework, however, is complicated by two factors: (i) ascertainment of families by case–control selection of probands, and (ii) missing genotype information for relatives. Li et al. (1998), Hsu et al. (1999), and Shih and Chatterjee (2002) developed different methods for analyzing familial data that can handle case–control ascertainment of probands; all of these methods, however, assume covariate information is available for both case–control probands and their relatives. In the following, we extend the method of Shih and Chatterjee (2002) to deal with missing genotype data in the relatives.

2.3 *Likelihoods*

The likelihood of the data under case–control sampling and with missing genotypes of the relatives can be written as (Gail et al., 1999a, 1999b; Chatterjee and Wacholder, 2001)

$$\begin{aligned}L &= \prod_{i=1}^{N} \Pr\big(\boldsymbol{Y}_i^R, G_i^P \,\big|\, Y_{i0}\big)\\ &= \prod_{i=1}^{N} \Pr\big(\boldsymbol{Y}_i^R \,\big|\, G_i^P, Y_{i0}\big) \times \prod_{i=1}^{N} \Pr\big(G_i^P \,\big|\, Y_{i0}\big)\\ &= L_{KC} \times L_{CC}.\end{aligned} \tag{4}$$

In (4), L_{KC} can be viewed as an ascertainment corrected likelihood for the "kin–cohort" data for the disease incidence data of the relatives and L_{CC} can be viewed as the retrospective case–control likelihood for the genotype data of the probands given their own disease history and age information. We observe that L_{CC} conditions on both the disease status and the age information for the probands. If cases and controls are selected randomly from the population of diseased and nondiseased subjects, an alternative likelihood for the proband's data could be formed by conditioning only on the disease status (Δ_i^P) of the probands. There are, however, two distinct disadvantages with such a likelihood. First, computation for such a likelihood could be complex as it would require modeling the censoring distribution for the probands. Second, such a likelihood will not be able to handle frequency-matched case–control studies where subjects are sampled conditional on both disease status and age-at-onset information. In contrast, the likelihood L_{CC}, by conditioning on both T^P and Δ^P, avoids modeling the censoring distribution and retains the ability to handle both matched and unmatched case–control studies.

Now we consider computation of the likelihood L (4) in terms of the model parameters defined in Section 2.2. First, to compute L_{CC}, we write

$$\begin{aligned}\Pr\big(G_i^P \,\big|\, Y_i^P\big) &= \frac{\Pr\big(Y_i^P \,\big|\, G_i^P\big)\Pr_f\big(G_i^P\big)}{\sum_{g\in\{0,1\}} \Pr\big(Y_i^P \,\big|\, G_i^P = g\big)\Pr_f\big(G_i^P = g\big)}\\ &= \frac{\lambda_{G_i^P}\big(T_i^P\big)^{\Delta^P} S_{G_i^P}\big(T_i^P\big)\Pr_f\big(G_i^P\big)}{\sum_{g\in\{0,1\}} \lambda_g\big(T_i^P\big)^{\Delta^P} S_g\big(T_i^P\big)\Pr_f\big(G_i^P = g\big)}.\end{aligned} \tag{5}$$

Above, the implicit assumption is that sampling of the probands is random conditional on their disease status Y^P. This assumption has important implications for prevalent case–control studies where cases may be sampled long after the incidence of the disease. If a disease is fatal and survival after disease incidence is related to the gene under study, the cases who survive to be recruited in the study would not be a representative sample of the underlying population of cases. In principle, it is possible to consider a more general form of the likelihood that can account for such survival bias (Gail et al., 1999a, 1999b; Gail and Chatterjee, 2004); computationally, however, such a generalization could be challenging.

Next, we consider computation of L_{KC}. We write,

$$\begin{aligned}
&\Pr\left(\boldsymbol{Y}_i^R \,\middle|\, G_i^P, Y_i^P\right) \\
&= \frac{\Pr\left(\boldsymbol{Y}_i^R, Y_i^P \,\middle|\, G_i^P\right)}{\Pr\left(Y_i^P \,\middle|\, G_i^P\right)} \\
&= \frac{\sum_{\boldsymbol{g}\in\{0,1\}^{M_i}} \Pr\left(\boldsymbol{Y}_i^R, Y_i^P \,\middle|\, \boldsymbol{G}_i = \boldsymbol{g}, G_i^P\right) \Pr_f\left(\boldsymbol{G}_i = \boldsymbol{g} \,\middle|\, G_i^P\right)}{\Pr\left(Y_i^P \,\middle|\, G_i^P\right)},
\end{aligned} \tag{6}$$

where $\sum_{\boldsymbol{g}\in\{0,1\}^{M_i}}$ represents the sum over all possible configurations of genotypes of the M_i relatives and $\Pr_f(\boldsymbol{G}_i = \boldsymbol{g} \mid G_i^P)$ denotes the probability of the particular genotype configuration $\boldsymbol{g}$ given the genotype of the proband, computed under allele frequency f. Following techniques for likelihood calculation in multivariate analysis, $\Pr(\boldsymbol{Y}_i^R,\ Y_i^P \mid \boldsymbol{G}_i = \boldsymbol{g},\ G_i^P)$ in formula (6) can be expressed in terms of the multivariate survivor function specified in equation (2) and its derivatives of various orders. In particular, for Clayton's copula model, the formula for $\Pr(Y_i^R,\ Y_i^P \mid \boldsymbol{G}_i = \boldsymbol{g},\ G_i^P)$ has been described in detail in formula (4) of Shih and Chatterjee (2002).

Chatterjee and Wacholder (2001) introduced a variant of the kin–cohort likelihood (L_{KC}) that has both computational and robustness advantages. In this approach, which they originally termed the "marginal likelihood," the family corresponding to a proband and his/her M relatives are broken into M relative-proband doublets and each such doublet is then treated independent of the others, ignoring possible dependence between doublets from the same family. More specifically, the "marginal likelihood," which more appropriately can be called the "composite likelihood," is written as

$$\begin{aligned}
&CL_{KC} \\
&= \prod_{i=1}^{N}\prod_{j=1}^{M_i} \Pr\left(Y_{ij}^R \,\middle|\, G_i^P, Y_i^P\right) \\
&= \prod_{i=1}^{N}\prod_{j=1}^{M_i} \frac{\sum_g \Pr\left(Y_{ij}^R, Y_i^P \,\middle|\, G_{ij} = g, G_i^P\right) \Pr_f\left(G_{ij} = g \,\middle|\, G_i^P\right)}{\Pr\left(Y_i^P \,\middle|\, G_i^P\right)},
\end{aligned} \tag{7}$$

where the bivariate probabilities $\Pr(Y_{ij}^R,\ Y_i^P \mid G_{ij}^R,\ G_i^P)$ can be computed as a special case of the computation for the multivariate survival probabilities that we discussed above. Moreover, the conditional probability of the gene-carrier status of the individual relative given that of the proband, $\Pr(G_{ij}^R \mid G_i^P)$, can be computed in terms of allele frequency (f) assuming standard Mendelian inheritance. Since its computation involves only pairs of individuals at a time, the complexity of the marginal likelihood, unlike that of the full likelihood, does not increase with family size. Moreover, marginal likelihood can be more robust to model misspecification for the multivariate distribution, as validity of the method relies only on the correct specification of the bivariate distribution for the relative-proband doublets.

2.4 *Estimation*

Direct maximization of the likelihood $L = L_{KC} \times L_{CC}$ or the composite likelihood $CL = CL_{KC} \times L_{CC}$ jointly with respect to the Euclidian parameters $\boldsymbol{\gamma} = (\boldsymbol{\beta},\ \theta,\ f)$ and nonparametric baseline hazard function $\lambda_0(t)$ is computationally difficult and numerically unstable. We developed an estimation procedure that is computationally tractable and yet retains statistical efficiency by identifying different parts of the likelihood which are most informative about the different parameters of interest.

We first observe that $\lambda_0(t)$ is involved both in the kin–cohort likelihood L_{KC} (or CL_{KC}) and in the case–control likelihood L_{CC}. Intuitively, however, it is clear that L_{CC}, due to conditioning on both disease status and age-at-onset information, cannot contain much information about the baseline risk of a disease. Thus, although in our modeling framework $\lambda_0(t)$ can be theoretically identifiable from L_{CC}, in practice L_{CC} would be expected to contain very little information about $\lambda_0(t)$. Moreover, from the functional form of L_{CC} given in equation (5) we observe that optimization of L_{CC} with respect to the nonparametric baseline hazard $\lambda_0(t)$ would be numerically challenging. Later we describe a simple and stable algorithm for estimation of $\lambda_0(t)$ for fixed value of $\boldsymbol{\gamma}$, which utilizes information from the kin–cohort likelihood CL_{KC}.

Next, consider the Euclidian parameters $\boldsymbol{\gamma} = (\theta,\ f,\ \beta)$. Clearly, the case–control likelihood L_{CC} does not involve the familial aggregation parameter θ. Although both L_{KC} and CL_{KC} involve the allele frequency parameter f, the kin–cohort likelihoods are not intrinsically very informative about f, as the relatives' genotype data are not directly observed. To avoid the related numerical problems, we propose to estimate f using only the case–control likelihood L_{CC}. Finally, we observe that both case–control and kin–cohort data are informative about hazard-ratio parameters ($\boldsymbol{\beta}$) and thus we propose to estimate them using the combined likelihood $L = L_{KC} \times L_{CC}$ or $CL = CL_{KC} \times L_{CC}$.

The final algorithm for implementing the estimation method iterates among the following steps:

- For fixed values of $\theta^{(s)}$, $f^{(s)}$, and $\lambda_0^{(s)}(t)$, maximize $CL_{KC} \times L_{CC}$ or $L_{KC} \times L_{CC}$ with respect to β to obtain $\beta^{(s+1)}$.
- For fixed values of $\beta^{(s+1)}$, $f^{(s)}$, and $\lambda_0^{(s)}(t)$, maximize CL_{KC} or L_{KC} with respect to θ to obtain $\theta^{(s+1)}$.
- For fixed values of $\beta^{(s+1)}$, $\theta^{(s+1)}$, and $\lambda_0^{(s)}(t)$, maximize L_{CC} with respect to f to obtain $f^{(s+1)}$.
- For fixed values of $\beta^{(s+1)}$, $f^{(s+1)}$, and $\theta^{(s+1)}$, use the following expectation-solving (ES) algorithm to estimate $\lambda^{(s+1)}$.

2.4.1 *ES algorithm for nonparametric estimation of $\lambda_0(t)$ for fixed $\boldsymbol{\gamma}$.*

To introduce this method, some counting process notation is useful. Define the processes $Y_{ij}^R(u) = I(T_{ij}^R \geq u)$, $N_{ij}^R(u) = I(T_{ij}^R \leq u)$, and $N_{++}^R(u) = \sum_{ij} N_{ij}^R(u)$. Based on results given is Hsu et al. (1999), we can write the hazard function of a relative given his/her own genotype (G^R) and the proband's disease status (T^P, Δ^P) and genotype (G^P) as

$$\begin{aligned}
&\lambda^R(t \mid T^P, \Delta^P, G^P, G^R) \\
&\quad = \lambda^R(t \mid T^P, \Delta^P = 0, G^P, G^R)\{\psi_\theta(u, v)\}^{\Delta_P},
\end{aligned} \tag{8}$$

where

$$\psi_\theta(u,v)=\frac{C_\theta(u,v)\partial C_\theta(u,v)/\partial u\partial v}{\{\partial C_\theta(u,v)/\partial u\}\{\partial C_\theta(u,v)/\partial v\}},$$

denotes the cross-ratio function for local measure of dependence between a pair of survival times (Clayton, 1978). In words, the cross-ratio function $\psi_\theta(u,\ v)$ evaluated at $u=S_{G^R}(t)$ and $v=S_{G^P}(T^P)$ can be interpreted as the "hazard-ratio" for a relative at time t associated with the case–control status ($\Delta^P=1$ vs. 0) of the index proband. Under Clayton's copula model (equation (3)), $\psi_\theta(u,\ v)=\theta$ is constant over time, with the values of $\theta=1$, $\theta>1$, and $\theta<1$ corresponding to independence, positive dependence, and negative independence, respectively. Shih and Chatterjee (2002) further showed that for copula models one can write

$$\begin{aligned}&\lambda^R(t\,|\,T^P,\Delta^P=0,G^P,G^R)\\&\quad=\lambda_0(t)e^{G^R\beta(t)}\phi_\theta\big\{S_{G^R}(t),S_{G^P}(T^P)\big\},\end{aligned}\tag{9}$$

where

$$\phi_\theta(u,v)=u\frac{\partial C_\theta(u,v)}{\partial u}\Big/C_\theta(u,v).$$

Based on formulas (8) and (9), which assume that the genotypes of the relatives are available, Shih and Chatterjee (2002) observed that $\lambda^R(t\,|\,T^P,\ \Delta^P,\ G^P,\ G^R)$ has a "PH" form with "time-dependent covariates," which, other than $\beta(t)$, depend on the survival times of probands and relatives only through the values of the associated marginal survival functions. Thus, they proposed estimating $\lambda_0(t)$ by iteratively fitting a Nelson–Aalen estimator that corresponds to solving the unbiased estimating equation,

$$d\Lambda_0(t)S^{(0)}(\boldsymbol{\beta},\theta,\Lambda_0,t)=dN_{++}(t),\tag{10}$$

where

$$\begin{aligned}S^{(0)}(\boldsymbol{\beta},\theta,\Lambda_0,t)=&\sum_{i=1}^{n}\sum_{j=1}^{m_i}I\big(T_{ij}^R\geq t\big)\\&\times e^{G_{ij}^R\beta(t)}H_{G_{ij}^R,G_i^P,T_i^P,\Delta_i^P}(t;\boldsymbol{\beta},\theta,\Lambda_0),\end{aligned}$$

with

$$\begin{aligned}H_{G^R,G^P,T^P,\Delta^P}(t;\boldsymbol{\beta},\theta,\Lambda_0)=&\,\phi_\theta\big\{S_{G^R}(t),S_{G^P}(T^P)\big\}\\&\times\psi_\theta\big\{S_{G^R}(t),S_{G^P}(T^P)\big\}^{\Delta^P}.\end{aligned}$$

In our context, the genotypes of the relatives (G_{ij}^R) are not known; thus the estimating equation (10) cannot be used directly. In this case, a natural way to obtain an unbiased estimating equation is to consider the conditional expectation of the estimating equation (10) given the observed data. In particular, we propose the estimating equation

$$d\Lambda_0(t)\widetilde{S}^{(0)}(\boldsymbol{\beta},\theta,\Lambda_0,t)=dN_{++}(t),\tag{11}$$

where

$$\begin{aligned}&\widetilde{S}^{(0)}(\boldsymbol{\beta},\theta,\Lambda_0,t)\\&=\sum_{i=1}^{m}\sum_{j=1}^{n_i}I\big(T_{ij}^R\geq t\big)\\&\quad\times E\Big\{e^{G_{ij}^R\beta(t)}H_{G_{ij}^R,G_i^P,T_i^P,\Delta_i^P}(t;\boldsymbol{\beta},\theta,\Lambda_0)\,\Big|\,T_{ij}^R,\Delta_{ij}^R,T_i^P,\Delta_i^P,G_i^P\Big\}.\end{aligned}\tag{12}$$

The unbiasedness of the estimating equation (11) follows from a standard conditional expectation argument showing that $ES^{(0)}(\boldsymbol{\beta},\theta,\Lambda_0,t)=E\widetilde{S^{(0)}}(\boldsymbol{\beta},\theta,\Lambda_0,t)$. The forms of equations (11) and (12) suggest an ES algorithm, where, at the $(s+1)$th iteration, the E-step evaluates the conditional expectation in formula (11) with respect to the conditional distribution

$$\begin{aligned}&\Pr_{\boldsymbol{\gamma},\Lambda_0^{(s)}}\big\{G_{ij}^R=g\,\big|\,T_{ij}^R,\Delta_{ij}^R,T_i^P,\Delta_i^P,G_j^P\big\}\\&=\frac{\Pr_{\boldsymbol{\beta},\theta,\Lambda_0^{(s)}}\big(T_{ij}^R,\Delta_{ij}^R,T_{ij}^P,\Delta_{ij}^P\,\big|\,G_{ij}^R=g,G_{ij}^P\big)\Pr_f\big(G_{ij}^R=1\,\big|\,G_{ij}^P\big)}{\sum_{g'}\Pr_{\boldsymbol{\beta},\theta,\Lambda_0^{(s)}}\big(T_{ij}^R,\Delta_{ij}^R,T_{ij}^P,\Delta_{ij}^P\,\big|\,G_{ij}^R=g',G_{ij}^P\big)\Pr_f\big(G_{ij}^R=g'\,\big|\,G_{ij}^P\big)},\end{aligned}$$

and the S-step updates $\lambda_0(t)$ using the *closed-form* formula

$$\lambda_0^{(s+1)}\big(t_{(i)}\big)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\delta_{ij}^R I\big(T_{ij}^R=t_{(i)}\big)}{\widetilde{S}_0\big(u;\boldsymbol{\beta},\theta,\Lambda_0^{(s)}\big)}.$$

In this estimation strategy, information on the relative risk parameters β is derived from both kin–cohort and case–control data. The underlying assumption is that the relative risk associated with the genetic variant is homogeneous in the population of the relatives and in the population from which the case–control sample has been selected. The assumption can be violated in several ways. The association between the genetic variant and the disease may be distorted in the case–control sample if cases under study are prevalent and the survival after disease incidence is related to the genetic variant under study. The disease history data of the relatives, in contrast, are not affected by such survival bias, as data on the relatives are collected through interview of the probands. Nonhomogeneity of relative risk can also occur if the case–control probands are selected from a special population. Investigators in the National Cancer Institute, for example, are currently conducting a nested case–control study of breast cancer within a cohort of women who work as radiation technologists. The disease history data of the relatives of the selected case–control subjects give us an opportunity for kin–cohort analysis. We observe that members of the radiation technologist cohort are exposed to certain background levels of radiation. The relatives of the cases and controls, however, generally do not work as radiation technologists and hence do not have such background exposure. The effect of the genetic variant on the populations with and without exposure to background radiation may be different if the effect of the genetic variant is modified by radiation exposure.

A comparison of two analyses can be performed to investigate the possibility of the nonhomogeneity of relative risk estimates. In the first step of the algorithm described above,

one can estimate $\boldsymbol{\beta}$ by maximizing only the kin–cohort likelihood (L_{KC} or CL_{KC}) and by maximizing only the case–control likelihood (L_{CC}). Iterating among steps 1–5 one can then obtain two estimates of the relative risk parameters, say $\hat{\beta}_{KC}$ and $\hat{\beta}_{CC}$, which pertain to the underlying populations for the kin–cohort and the case–control data, respectively. The estimates can be compared, either formally or informally, to examine whether the assumption of homogeneous relative risk is realistic.

2.5 *Asymptotic Theory and Variance Estimation*

For a parametric model, such as a fixed-knot piecewise exponential model for the baseline hazard function, the consistency of the various estimation methods we described follows from standard estimating equation theory (Godambe, 1991). For a nonparametric baseline hazard, which allows a knot at each of the observed event times, a similar consistency result can be expected to hold, although a rigorous proof is not yet available. Simulation studies reported in Section 4 show that the proposed semiparametric estimation methods are consistent both for the finite dimensional parameters $\boldsymbol{\beta}$, θ, f and for the nonparametric baseline hazard function $\lambda_0(t)$.

In principle, one can use estimating-equation-based variance estimators such as the so-called robust-sandwich method that is widely used in the generalized estimating equation (GEE) literature (Liang, Zeger, and Quaquish, 1992). Although these methods work well for parametric models, their performance in semiparametric settings has not been thoroughly studied. In our data application, we used a bootstrap-based resampling method (Efron and Tibshirani, 1998) that is known to perform well for both parametric and nonparametric models. To account for possible familial correlation between the relatives of the same proband, we use families as the bootstrap sampling units. If there are $N_1 + N_0$ unique families corresponding to $N_1 + N_0$ case–control probands in the study, in each bootstrap sample we draw N_1 and N_0 families with replacement from the N_1 and N_0 families of case and control probands, respectively. Once a bootstrap sample of the families is chosen, the proposed method is used to obtain bootstrap estimates of the parameters. The empirical percentiles for the bootstrap estimates over different bootstrap samples are used to define the confidence intervals for the parameter estimates.

3. Estimating Penetrance from the Case-Only Design

Once a genetic mutation has been associated with a disease, estimation of the cumulative risk of the disease (penetrance) for subjects who carry the genetic mutation can be important both for estimating the impact on public health and genetic counseling. In recent years, a number of studies have used kin–cohort data from a sample of genotyped cases to estimate the risk of various cancers associated with highly penetrant rare genes such as BRCA1 and BRCA2. Use of case probands is convenient, because recruitment of population-based controls who are willing to provide samples can be expensive and subject to selection bias. Moreover, mutations in major genes such as BRCA1 and BRCA2 are very rare in the general population; population-based controls will typically have very few carriers. The cases and their relatives, on the other hand, have a much higher frequency of the mutation carriers. Thus, the kin–cohort analysis of family history data from a series of genotyped cases yields a relatively quick and inexpensive estimate of penetrance for a rare genetic mutation.

The validity of penetrance estimates based on kin–cohort data from case probands only has been questioned. Our earlier work (Wacholder, 1998; Gail et al., 1999a, 1999b; Chatterjee and Wacholder, 2001) recognized that penetrance estimates for mutation carriers will be upwardly biased in the presence of residual familial aggregation if the analytic methods assume that disease risk depends only on mutation status. Begg (2002) elaborated on the possibility of such bias as a consequence of size-biased sampling theory (Patil and Rao, 1978). He reviewed eight published studies on the risk of breast cancer associated with BRCA1/2 mutations and noted that estimates of penetrance from studies that used the family history of case probands tend to be higher than those from a study that was not susceptible to such bias. The authors concluded that future methodological research is needed to correct for such bias in kin–cohort studies. Whittemore and Gong (2003) and Gong and Whittemore (2003), on the other hand, used both the empirical data reviewed by Begg and simulation studies to show that the degree of bias due to use of case probands only is small relative to the standard error of the estimates.

For the analysis of kin–cohort data from case probands only, we propose the use of the ascertainment-corrected kin–cohort likelihood L_{KC} (see formula (6)). Modeling and estimation of the residual familial aggregation parameter θ is key to adjustment for ascertainment in our approach. In Section 2.3, we described how the residual familial aggregation parameter θ can be estimated from kin–cohort data, using either the full likelihood L_{KC} or the composite likelihood CL_{KC}. In the composite-likelihood approach, which is based on relative-proband doublets, the information on θ is extracted from the comparison of disease incidence among the relatives of cases with that among the relatives of controls (see formula (8) and subsequent discussion). Obviously, such comparisons are not possible when probands consist of cases only; thus the composite likelihood CL_{KC} intrinsically contains very little information on θ for case-only designs. This problem with composite likelihood is manifested in numerical instability in optimization when estimating θ. In the joint likelihood L_{KC}, on the other hand, the information on θ is derived not only from comparison of disease incidence in relatives of cases and controls, but also from familial aggregation of the disease among the relatives of the same proband. Thus, even if the probands consist of cases only, reliable estimates of θ can be obtained based on joint disease incidence data of the relatives from the same proband, which in turn can be used to correct penetrance estimation for ascertainment.

We found that estimation of f internally is numerically unstable when genotype data are not available for controls. Thus, for analysis of data from case probands only we will assume that the allele frequency parameter f is known or can be estimated externally, and the third step in the iterative scheme in Section 2.4 is omitted. Fortunately, the kin–cohort analysis itself is not very sensitive to the value of f, as the conditional probabilities $\Pr_f(\boldsymbol{g} \mid g^P)$ involved in the likelihood L_{KC} do not vary much with f.

Although many recent analyses are comparable to use of L_{KC} alone, the likelihood for the proband's data, L_{CC}, can be informative for estimation of relative risk parameters even if the probands consist of cases only. Thus, we considered an alternative analysis for the case-only design in which we estimated β by maximizing $L_{KC} \times L_{CC}$. This approach, however, would be more sensitive to misspecification of *f*, as the likelihood L_{CC} depends on the unconditional probabilities $\Pr_f(g^P)$, which strongly depend on *f*.

4. Simulation Studies

4.1 *Performance Assessment*

4.1.1 *Case–control design.* We studied performance of the proposed methods through simulated studies of breast cancer involving a dominant gene and family structure consisting of proband, sister, and mother. We considered two simulation scenarios, one involving a rare generic variant ($f = 0.01$) with large effect (hazard ratio = 5) and the other involving a more common variant ($f = 0.2$) with a modest effect (hazard ratio = 2.0). Given genotypes of the three relatives, we assumed their joint time-to-cancer incidence follows Clayton's copula model with the marginal distributions for the three relatives specified by Cox's proportional hazard model. In particular, we assumed the marginal risk for a noncarrier ($G = 0$) relative follows a Weibull distribution with shape and scale parameters chosen so that cumulative risk until age 50 and 70 years is 5% and 13%, respectively. The marginal risk for a carrier ($G = 1$) was then specified so that the hazard for a carrier was proportional to the hazard for a noncarrier, with proportionality constant (hazard ratio) being 5 and 2 for the rare and the common variants, respectively. We chose the association parameter θ in the copula model to be 2.0 which approximately corresponds to a 2-fold higher risk of breast cancer among relatives of breast cancer cases compared to relatives of breast cancer controls in our simulation setup.

Based on the above model, we first generated data for a random sample of families of three individuals: (proband, sister, mother). We generate the genotype ($G = 1$ or 0) for three relatives based on Mendelian probabilities. Given genotypes of the relatives, we generated time-to-cancer incidence (T) for the relatives from the trivariate copula distribution by first generating data for the proband from the appropriate marginal distribution, then for the sister given the proband, and finally for the mother given the sister and the proband from appropriate conditional distributions. Data from the marginal and the conditional distributions were generated by appropriate inverse distribution function transformations of uniform distributions. After generating time-to-cancer incidence, we generated censoring/current age (C) for probands, sisters, and mothers from independent normal distributions with means 50, 50, and 70, respectively, and with a common variance of 10. Following standard convention, we assume the observed data consist of whether the relatives had cancer ($T \leq C$, $\Delta = 1$) or not ($T > C$, $\Delta = 0$), the age at onset (T) for relatives who had cancer ($\Delta = 1$), and the current/censoring age (C) for relatives who were cancer free ($\Delta = 0$). Following the above scheme, we randomly sampled a large number of triplets, which in turn were used as the source population from which 2000 triplets with case probands ($\Delta = 1$) and 2000 triplets with control probands ($\Delta = 0$) were selected.

We analyzed each simulated data set using two approaches: in one $\boldsymbol{\beta}$ and θ were estimated using the composite likelihood $CL_{KC} \times L_{CC}$ and in the other $\boldsymbol{\beta}$ and θ were estimated using the joint likelihood $L_{KC} \times L_{CC}$. In both approaches, f was estimated using L_{CC} and the baseline hazard function was estimated using the nonparametric ES algorithm described in Section 2.4. We assumed genotype data are available only for probands, but not for relatives. To examine the performance of the proposed method in estimating age-specific hazard ratios, we analyzed each data set using a piecewise proportional hazard model that allows estimation of separate hazard ratios for age intervals: <50, 50–60, 60–70, and >70 years. Compared to previous methods (Gail et al., 1999a, 1999b; Chatterjee and Wacholder, 2001; Moore et al., 2001), one novel aspect of the current methodology is that it accounts for residual familial correlation of the disease that cannot be explained by the gene under study. To examine the benefit of such extension, we implemented both the composite and joint likelihood approaches with θ fixed at 1.0, which corresponds to assumption of no residual familial correlation, and compared the results with those in which we estimated θ from the data. Table 1 shows simulated averages and standard errors for estimates of hazard ratio parameters (β), residual correlation parameters (θ), and cumulative risks up to age 50 and 70 years for noncarriers ($F_0(50)$ and $F_0(70)$) and those for carriers ($F_1(50)$ and $F_1(70)$).

We first observe that when $\theta = 1$ is assumed, the cumulative risks for noncarriers are substantially overestimated (Table 1). Intuitively, this is expected. When $\theta = 1$ or no residual familial aggregation has been assumed, the kin–cohort likelihood of relatives' disease incidence data becomes free of the ascertainment mechanism that is defined by the case–control status (Y_{i0}) of the probands. An estimate of $F_0(t)$ based on CL_{KC} that is not adjusted for case–control ascertainment overestimates the baseline risk for the general population, as relatives of a case-enriched sample of probands are expected to have higher disease incidence than that of the general population. The hazard ratio parameters are also overestimated; the magnitude of the bias, however, is generally smaller when β is estimated using $CL_{KC} \times L_{CC}$ than using $L_{KC} \times L_{CC}$. The robustness of the composite likelihood method can be attributed to the fact that CL_{KC}, unlike L_{KC}, is only affected by correlation between the relatives and the proband, but not by the correlation between relatives of the same proband. Finally, we observe that when $\theta = 1$ is assumed, the upward biases in $F_0(t)$ create an upward bias in the estimation of $F_1(t)$. In contrast, when θ is estimated from the data, which in turn is then used to account for case–control ascertainment, the biases in estimates of the cumulative risk as well as hazard ratio parameters disappear while the corresponding standard errors only increase slightly. The precision of estimates of the hazard ratio parameters was very similar for the joint likelihood $L_{KC} \times L_{CC}$ and the composite likelihood $CL_{KC} \times L_{CC}$. For estimation of θ, however, a slight gain in efficiency is observed for the joint likelihood.

Table 2 shows the relative informativeness of kin–cohort and case–control data for estimation of the hazard ratio parameters (β), computed as the inverse variance for estimates of β using L_{KC}- (or CL_{KC}-) only and L_{CC}- only, respectively, as a ratio to that when β is estimated using $L_{KC} \times L_{CC}$ (or

Table 1

Case–control studies: simulation study to assess bias due to the assumption of no residual familial aggregation ($\theta = 1$) and performance of the proposed methodology that estimates familial aggregation from the data assuming Clayton's copula model

	High-risk rare gene			Moderate-risk common gene		
Method used	True parameter	$\theta = 1$ Mean (SE)	Estimated θ Mean (SE)	True parameter	$\theta = 1$ Mean (SE)	Estimated θ Mean (SE)
$CL_{KC} \times L_{CC}$	$\exp(\beta_1) = 5$	5.02 (0.63)	4.96 (0.63)	$\exp(\beta_1) = 2$	2.03 (0.14)	1.99 (0.13)
	$\exp(\beta_2) = 5$	5.29 (1.05)	4.92 (0.99)	$\exp(\beta_2) = 2$	2.07 (0.22)	1.99 (0.21)
	$\exp(\beta_3) = 5$	5.58 (1.44)	4.91 (1.29)	$\exp(\beta_3) = 2$	2.13 (0.31)	1.99 (0.28)
	$\exp(\beta_4) = 5$	5.86 (3.10)	4.82 (2.66)	$\exp(\beta_4) = 2$	2.18 (0.48)	1.96 (0.42)
	$\theta = 2$	NA	1.99 (0.13)	$\theta = 2$	NA	1.99 (0.12)
	$f = 0.01$	0.01 (0.001)	0.01 (0.001)	$f = 0.2$	0.20 (0.006)	0.20 (0.006)
	$F_0(50) = 0.05$	0.06 (0.003)	0.05 (0.003)	$F_0(50) = 0.05$	0.06 (0.004)	0.05 (0.003)
	$F_0(70) = 0.13$	0.17 (0.010)	0.13 (0.007)	$F_0(70) = 0.13$	0.17 (0.008)	0.13 (0.008)
	$F_1(50) = 0.20$	0.28 (0.030)	0.21 (0.023)	$F_1(50) = 0.09$	0.12 (0.006)	0.09 (0.005)
	$F_1(70) = 0.50$	0.63 (0.050)	0.49 (0.050)	$F_1(70) = 0.24$	0.32 (0.014)	0.24 (0.013)
$L_{KC} \times L_{CC}$	$\exp(\beta_1) = 5$	5.30 (0.66)	4.96 (0.63)	$\exp(\beta_1) = 2$	2.13 (0.15)	1.99 (0.13)
	$\exp(\beta_2) = 5$	5.67 (1.14)	4.92 (0.99)	$\exp(\beta_2) = 2$	2.21 (0.24)	1.99 (0.21)
	$\exp(\beta_3) = 5$	6.12 (1.59)	4.92 (1.30)	$\exp(\beta_3) = 2$	2.35 (0.36)	1.99 (0.28)
	$\exp(\beta_4) = 5$	6.74 (3.45)	4.85 (2.77)	$\exp(\beta_4) = 2$	2.60 (0.62)	1.96 (0.43)
	$\theta = 2$	NA	2.00 (0.12)	$\theta = 2$	NA	2.00 (0.11)
	$f = 0.01$	0.01 (0.001)	0.01 (0.001)	$f = 0.2$	0.19 (0.006)	0.20 (0.006)
	$F_0(50) = 0.05$	0.06 (0.003)	0.05 (0.003)	$F_0(50) = 0.05$	0.06 (0.004)	0.05 (0.003)
	$F_0(70) = 0.13$	0.17 (0.006)	0.13 (0.007)	$F_0(70) = 0.13$	0.17 (0.008)	0.13 (0.008)
	$F_1(50) = 0.20$	0.29 (0.029)	0.21 (0.023)	$F_1(50) = 0.09$	0.13 (0.006)	0.09 (0.005)
	$F_1(70) = 0.50$	0.66 (0.051)	0.49 (0.050)	$F_1(70) = 0.24$	0.33 (0.015)	0.24 (0.013)

$CL_{KC} \times L_{CC}$). Kin–cohort data from the relatives add substantial information on β in addition to the traditional case–control analysis for all age groups, even though relatives are not genotyped. The gain is more substantial for the rare variant and for the older age groups for which the case–control analysis lacks power due to small numbers of variants. The added value of kin–cohort analysis for relative risk estimation for common cancers, such as breast cancer, was also observed by Saunders and Begg (2003).

4.1.2 *Case-only design.* In Section 3, we described how the ascertainment-corrected kin–cohort likelihood L_{KC} or the combined kin–cohort and proband's likelihood $L_{KC} \times L_{CC}$ can be used to analyze the disease incidence data of the relatives of a sample of genotyped cases. We studied the performance of this approach using the same simulation setup as above except that now we assumed data were available only from the sample of case probands. Table 3 shows simulation results when data were analyzed assuming the correctly specified Clayton's model.

For the study of a highly penetrant rare gene using L_{KC} alone, assumption of no residual familial aggregation ($\theta = 1$) leads to slight underestimation of the hazard ratio parameters and substantial, but not severe, overestimation of the penetrance parameters $F_1(50)$ and $F_1(70)$. In the same scenario, when the combined likelihood $L_{KC} \times L_{CC}$ was used, the assumption of $\theta = 1$ leads to overestimation of the hazard ratio parameters and larger bias for the cumulative risk parameters. For the low penetrant common gene, on the other hand, the assumption of $\theta = 1$ causes very little bias for estimation of the hazard ratio parameters, but results in severe overestimation of the penetrance parameters $F_1(50)$ and $F_1(70)$. When θ is estimated from the data assuming the correct model for residual familial correlation, the bias in estimates of all the parameters becomes much smaller. Use of $L_{KC} \times L_{CC}$ instead

Table 2

Case–control studies: relative efficiency for log of hazard ratios from kin–cohort and case–control analysis

	High-risk rare gene				Moderate-risk common gene			
	ARE compared to $CL_{KC} \times L_{CC}$		ARE compared to $L_{KC} \times L_{CC}$		ARE compared to $CL_{KC} \times L_{CC}$		ARE compared to $CL_{KC} \times L_{CC}$	
Parameter	CL_{KC}	L_{CC}	L_{KC}	L_{CC}	CL_{KC}	L_{CC}	L_{KC}	L_{CC}
β_1	0.39	0.55	0.38	0.54	0.12	0.87	0.12	0.86
β_2	0.39	0.56	0.40	0.53	0.16	0.80	0.16	0.80
β_3	0.48	0.39	0.49	0.39	0.27	0.75	0.26	0.75
β_4	0.67	0.49	0.58	0.44	0.49	0.43	0.49	0.44

Table 3

Case–only studies: simulation study to assess bias due to the assumption of no residual familial aggregation ($\theta = 1$) and performance of the proposed methodology that estimates familial aggregation from the data assuming Clayton's copula model

	High-risk rare gene			Moderate-risk common gene		
Method used	True parameter	$\theta = 1$ Mean (SE)	Estimated θ Mean (SE)	True parameter	$\theta = 1$ Mean (SE)	Estimated θ Mean (SE)
$CL_{KC} \times L_{CC}$	$\exp(\beta_1) = 5$	4.40 (0.84)	4.61 (1.34)	$\exp(\beta_1) = 2$	1.98 (0.42)	1.99 (0.42)
	$\exp(\beta_2) = 5$	4.31 (1.37)	4.54 (1.68)	$\exp(\beta_2) = 2$	2.03 (0.64)	2.06 (0.66)
	$\exp(\beta_3) = 5$	4.31 (1.81)	4.58 (2.12)	$\exp(\beta_3) = 2$	2.04 (0.70)	2.10 (0.73)
	$\exp(\beta_4) = 5$	4.16 (5.37)	4.53 (5.76)	$\exp(\beta_4) = 2$	1.86 (0.75)	1.89 (0.81)
	$\theta = 2$	NA	1.93 (0.52)	$\theta = 2$	NA	2.07 (0.60)
	$F_0(50) = 0.05$	0.08 (0.005)	0.05 (0.010)	$F_0(50) = 0.05$	0.08 (0.010)	0.05 (0.020)
	$F_0(70) = 0.13$	0.23 (0.010)	0.14 (0.030)	$F_0(70) = 0.13$	0.23 (0.019)	0.14 (0.050)
	$F_1(50) = 0.20$	0.32 (0.045)	0.20 (0.060)	$F_1(50) = 0.09$	0.16 (0.016)	0.09 (0.030)
	$F_1(70) = 0.50$	0.67 (0.069)	0.47 (0.120)	$F_1(70) = 0.24$	0.40 (0.030)	0.26 (0.070)
$L_{KC} \times L_{CC}$	$\exp(\beta_1) = 5$	4.95 (0.50)	4.99 (0.50)	$\exp(\beta_1) = 2$	2.03 (0.11)	2.01 (0.11)
	$\exp(\beta_2) = 5$	5.42 (0.97)	4.93 (1.00)	$\exp(\beta_2) = 2$	2.09 (0.21)	2.00 (0.20)
	$\exp(\beta_3) = 5$	5.82 (1.56)	4.96 (1.42)	$\exp(\beta_3) = 2$	2.17 (0.32)	2.01 (0.29)
	$\exp(\beta_4) = 5$	6.39 (3.98)	4.99 (3.31)	$\exp(\beta_4) = 2$	2.16 (0.53)	1.96 (0.47)
	$\theta = 2$	NA	1.99 (0.54)	$\theta = 2$	NA	2.07 (0.59)
	$F_0(50) = 0.05$	0.08 (0.005)	0.05 (0.011)	$F_0(50) = 0.05$	0.08 (0.005)	0.05 (0.011)
	$F_0(70) = 0.13$	0.23 (0.009)	0.13 (0.029)	$F_0(70) = 0.13$	0.22 (0.010)	0.13 (0.029)
	$F_1(50) = 0.20$	0.35 (0.028)	0.22 (0.051)	$F_1(50) = 0.09$	0.16 (0.008)	0.09 (0.022)
	$F_1(70) = 0.50$	0.75 (0.044)	0.51 (0.101)	$F_1(70) = 0.24$	0.41 (0.017)	0.24 (0.055)

of L_{KC} alone yields dramatically more precise estimates of relative risk parameters as well as more precise estimates of cumulative risks.

4.2 *Study of Robustness*

We studied the performance of the proposed methodology when the model for residual familial correlation is misspecified. In particular, we generated data using the same simulation scheme as above except that time-to-disease incidence for subjects in a family was generated from Frank's copula model instead of Clayton's model. We fixed the association parameter (θ) for Frank's model so that it corresponded overall to a 2-fold higher incidence of breast cancer among relatives of breast cancer cases than among relatives of breast cancer controls. Different copula models correspond to different patterns of age-specific familial aggregation. Clayton's model, for example, assumes risk is similar for relatives of early and late onset cases. Frank's model, in contrast, assumes moderately higher risk for relatives of early onset cases. Table 4 shows the bias and standard error of the estimates of different parameters of interest under the case–control design when the simulated data were analyzed assuming the incorrect Clayton model. We observe that bias due to the misspecified correlation model was generally quite small for both hazard ratio and cumulative risk parameters.

The simulation study revealed some intrinsic identifiability problems for analysis of data from the case-only design with a misspecified association model. For a large fraction of simulated data, the algorithm did not converge. Inspection of intermediate steps suggested that joint estimation of the baseline hazard $\lambda_0(t)$ and correlation parameter θ is unstable, although for a fixed value of one parameter the other parameter can be estimated quite reliably. In the case-only data, information on θ comes only from the joint incidence status of the relatives. Estimation of θ in this approach requires accounting for the marginal hazard for the individual relatives. However, estimation of marginal hazards using the relatives of cases requires correcting for ascertainment based on the correlation parameter θ itself. Thus, the circularity of this estimation problem may cause θ and $\lambda_0(t)$ to be jointly ill identified unless the correlation model is correctly specified.

We also studied the identifiability problem with the case-only design assuming the known Weibull parametric form for $\lambda_0(t)$. In this situation, the algorithm converged for about 90% of the simulated data sets, although the estimates of θ were often unstable with very large values. Table 5 shows the bias and standard error of hazard ratio and absolute risk parameters obtained using the converged data sets, ignoring the problem that the estimates of θ were unstable for some of the converged data sets. We observe that all of the cumulative risk parameters are severely underestimated both for analyses based on L_{KC} and $L_{KC} \times L_{CC}$. Estimates of the relative risk parameters are also biased, the biases often being larger than when $\theta = 1$ was assumed (Table 3).

Finally, we also studied the robustness of the two likelihoods L_{KC} and $L_{KC} \times L_{CC}$ to misspecification of the allele frequency parameter f (results not shown). In brief, we simulated data from a low penetrant ($\beta = \log(2)$) common gene ($f = 0.2$), but analyzed the data assuming $f = 0.3$. The copula model for residual correlation was correctly specified. The Monte Carlo average estimates of the hazard ratio parameters in the four age groups were 2.23, 2.33, 2.34, and 1.98 when L_{KC} was used and were 1.09, 1.07, 1.09, and 1.15 when $L_{KC} \times L_{CC}$ was used. Thus, the analysis based on $L_{KC} \times L_{CC}$ is very sensitive to misspecification of f.

Table 4
Simulation study to assess robustness of the case–control design against model misspecification of familial correlation. Data are generated from Frank's model, but are analyzed using Clayton's model.

	High-risk rare gene		Moderate-risk common gene	
Method used	True parameter	Mean (SE)	True parameter	Mean (SE)
$CL_{KC} \times L_{CC}$	$\exp(\beta_1) = 5$	4.78 (0.61)	$\exp(\beta_1) = 2$	1.99 (0.13)
	$\exp(\beta_2) = 5$	4.65 (0.96)	$\exp(\beta_2) = 2$	1.99 (0.21)
	$\exp(\beta_3) = 5$	4.45 (1.20)	$\exp(\beta_3) = 2$	1.97 (0.27)
	$\exp(\beta_4) = 5$	4.31 (2.23)	$\exp(\beta_4) = 2$	1.91 (0.43)
	$f = 0.01$	0.01 (0.001)	$f = 0.2$	0.20 (0.006)
	$F_0(50) = 0.05$	0.05 (0.003)	$F_0(50) = 0.05$	0.05 (0.003)
	$F_0(70) = 0.13$	0.13 (0.007)	$F_0(70) = 0.13$	0.13 (0.007)
	$F_1(50) = 0.20$	0.20 (0.023)	$F_1(50) = 0.09$	0.09 (0.005)
	$F_1(70) = 0.50$	0.47 (0.048)	$F_1(70) = 0.24$	0.24 (0.013)
$L_{KC} \times L_{CC}$	$\exp(\beta_1) = 5$	4.82 (0.62)	$\exp(\beta_1) = 2$	2.00 (0.13)
	$\exp(\beta_2) = 5$	4.69 (0.96)	$\exp(\beta_2) = 2$	2.00 (0.21)
	$\exp(\beta_3) = 5$	4.49 (1.22)	$\exp(\beta_3) = 2$	1.99 (0.28)
	$\exp(\beta_4) = 5$	4.38 (2.29)	$\exp(\beta_4) = 2$	1.94 (0.44)
	$f = 0.01$	0.01 (0.001)	$f = 0.2$	0.20 (0.006)
	$F_0(50) = 0.05$	0.05 (0.003)	$F_0(50) = 0.05$	0.05 (0.003)
	$F_0(70) = 0.13$	0.13 (0.007)	$F_0(70) = 0.13$	0.13 (0.008)
	$F_1(50) = 0.20$	0.20 (0.022)	$F_1(50) = 0.09$	0.09 (0.005)
	$F_1(70) = 0.50$	0.47 (0.048)	$F_1(70) = 0.24$	0.24 (0.013)

5. Example

We applied the proposed methodology to data from the Washington Ashkenazi Study (WAS) (1997). In this study, 5318 Ashkenazi Jewish volunteers living in the Washington, D.C. area were genotyped for three specific mutations in BRCA1 and BRCA2 genes and were interviewed for detailed personal and family history of various cancers. Struewing et al. (1997) estimated penetrance of breast, ovarian, and a number of other cancers associated with the BRCA1/2 mutations by analyzing cancer incidence data of the relatives using the original kin–cohort analytic approach of Wacholder et al. (1998). Chatterjee et al. (2001) and Moore et al. (2001) analyzed the breast cancer incidence data from the WAS relatives using various pseudo-likelihood approaches that corrected for

Table 5
Simulation study to assess robustness of the case-only design against model misspecification of familial correlation. Data are generated from Frank's model, but are analyzed using Clayton's model.

	High-risk rare gene		Moderate-risk common gene	
Method used	True parameter	Mean (SE)	True parameter	Mean (SE)
L_{KC}	$\exp(\beta_1) = 5$	5.20 (1.26)	$\exp(\beta_1) = 2$	2.28 (0.49)
	$\exp(\beta_2) = 5$	4.50 (1.44)	$\exp(\beta_2) = 2$	2.00 (0.41)
	$\exp(\beta_3) = 5$	3.96 (1.72)	$\exp(\beta_3) = 2$	1.80 (0.41)
	$\exp(\beta_4) = 5$	3.45 (2.66)	$\exp(\beta_4) = 2$	1.49 (0.45)
	$F_0(50) = 0.05$	0.02 (0.014)	$F_0(50) = 0.05$	0.02 (0.013)
	$F_0(70) = 0.13$	0.06 (0.042)	$F_0(70) = 0.13$	0.06 (0.038)
	$F_1(50) = 0.20$	0.10 (0.068)	$F_1(50) = 0.09$	0.05 (0.029)
	$F_1(70) = 0.50$	0.23 (0.149)	$F_1(70) = 0.24$	0.13 (0.072)
$L_{KC} \times L_{CC}$	$\exp(\beta_1) = 5$	4.83 (0.48)	$\exp(\beta_1) = 2$	2.01 (0.10)
	$\exp(\beta_2) = 5$	4.35 (0.79)	$\exp(\beta_2) = 2$	1.93 (0.16)
	$\exp(\beta_3) = 5$	3.88 (1.06)	$\exp(\beta_3) = 2$	1.82 (0.21)
	$\exp(\beta_4) = 5$	3.32 (1.90)	$\exp(\beta_4) = 2$	1.58 (0.29)
	$F_0(50) = 0.05$	0.02 (0.014)	$F_0(50) = 0.05$	0.02 (0.013)
	$F_0(70) = 0.13$	0.06 (0.041)	$F_0(70) = 0.13$	0.06 (0.038)
	$F_1(50) = 0.20$	0.09 (0.067)	$F_1(50) = 0.09$	0.04 (0.027)
	$F_1(70) = 0.50$	0.24 (0.153)	$F_1(70) = 0.24$	0.12 (0.071)

Table 6
Analysis of Washington Ashkenazi Study using Clayton's model for familial correlation

Parameter	CL_{KC} only Est (95% CI)	L_{CC} only Est (95% CI)	$CL_{KC} \times L_{CC}$ Est (95% CI)
HR[≤50] =	6.95 (3.73, 12.52)	8.58 (5.63, 12.57)	7.72 (4.99, 10.54)
HR[>50, ≤ 60] =	10.62 (4.57, 24.83)	6.22 (1.92, 11.65)	8.47 (4.61, 14.47)
HR[>60] =	3.17 (0.03, 19.18)	1.34 (0.003, 4.37)	2.20 (0.27, 7.09)
f =	–	0.01 (0.009, 0.014)	0.01 (0.009, 0.014)
θ =	1.35 (1.07, 1.68)	–	1.35 (1.07, 1.69)
$F_0(50)$ =	0.05 (0.045, 0.058)	–	0.05 (0.045, 0.057)
$F_0(70)$ =	0.14 (0.131, 0.155)	–	0.14 (0.132, 0.155)
$F_1(50)$ =	0.31 (0.188, 0.458)	–	0.33 (0.241, 0.418)
$F_1(70)$ =	0.64 (0.441, 0.923)	–	0.60 (0.486, 0.732)

the nonmonotonicity problem in the penetrance estimates obtained from the method of moment of approach of Wacholder et al. (1998).

In all of the previous analyses of WAS data, estimates of penetrance parameters utilized information only from the kin–cohort disease incidence data of the relatives, but not from the case–control data of the probands (volunteers). Moreover, in all of these analyses the possibility of bias due to presence of residual familial aggregation was ignored. Although in this study the volunteers were not selected based on case–control sampling, the effect of proband ascertainment was a potential concern, as it was observed that women with personal history of a breast cancer were more likely to volunteer than women who did not have such history. We applied the proposed composite-likelihood methodology to breast cancer data available from the female volunteers of this study. This novel analysis efficiently combines the family history (kin–cohort) and personal history (case–control) data of the volunteers, produces an estimate of residual familial aggregation, and accounts for the potential effect of differential participation of the cases and controls.

We first examined the homogeneity of the hazard ratio parameters between the population underlying the sample of volunteers and the population underlying the relatives of the volunteers. The first two columns of Table 6 show the hazard ratio estimates and 95% bootstrap confidence intervals (CI) based on CL_{KC}- only and L_{CC}- only. Both sets of estimates show similar patterns of age-specific risk from BRCA1/2 mutations and suggest much stronger effects of the mutations before the age of 60 years. Quantitatively, there are some differences between the individual point estimates; these differences, however, are well within the limits of uncertainty and hence are not statistically significant. Estimates of various parameters based on $CL_{KC} \times L_{CC}$ (kin–cohort and case–control likelihood combined) are shown in the third column of Table 6. We observe that the estimate of lifetime penetrance ($F_1(70)$) as 0.60 (95% CI: 0.49, 0.76) was very close to its estimates from all previous studies that did not account for ascertainment. Thus, it seems differential participation of the cases and controls in this study did not create bias in previous estimates of penetrance. Our analysis also shows the value of utilizing information on the hazard ratio parameters that is contained in the case–control likelihood. The estimate of penetrance based only on the kin–cohort likelihood of the relatives was 0.64, a number that was very similar to the overall estimate from the combined analysis, but had a much larger confidence interval.

We estimated the residual familial correlation parameter in Clayton's model to be 1.35 (95% CI: 1.07, 1.68), which corresponds to approximately 1.35-fold increased risk of breast cancers associated with a family history of the disease among noncarriers of the BRCA1/2 mutation. We note that in the composite-likelihood methodology the information on the correlation parameter is obtained from the relationship between the proband and the relatives. In contrast, in a previous study where we analyzed the correlation between the relatives of the same proband (Chatterjee et al., 2001), we had estimated about a 2-fold increased risk of breast cancer associated with family history of the disease among noncarriers. Similar strength of residual familial aggregation of breast cancer after accounting for the BRCA1/2 mutation was also reported by Claus et al. (1998) based on the Cancer and Steroid Hormone (CASH) case–control study. Because in the composite-likelihood approach the estimate of familial aggregation is essentially based on the comparison of family history between cases and controls, a possible explanation for the attenuated estimate of familial aggregation in this analysis is that controls with a family history of breast cancer were more likely to participate in this study, diminishing the true difference between family history of cases and controls.

6. Discussion

In summary, we have developed a methodology for analyzing case–control studies of genetic variants that collect detailed disease history information, but not genotypes, on relatives. We consider a modeling approach that is flexible, has a marginal parameter interpretation, and yet remains computationally tractable. We develop an estimation methodology that combines information on relative risk parameters from kin–cohort data of relatives and case–control data of participants. In addition, the method estimates the baseline risk and familial aggregation parameters using the kin–cohort data of the relatives. We also propose some modifications that can be used to analyze disease incidence data of relatives from a sample of genotyped cases only.

We have utilized copula models that assume the same correlation for different pairs of relatives. The proposed composite-likelihood approach can be easily generalized to allow for

different correlation (θ) parameters for different types of relative-proband relationships. The joint-likelihood approach, however, would require specifying a full joint distribution for all relatives in a family that can incorporate different pairwise correlation parameters. Such multivariate models are not currently available within the framework of copula modeling. Future research is merited for such extensions.

When case and control probands are available, the composite likelihood ($CL_{KC} \times L_{CC}$) or the joint likelihood ($L_{KC} \times L_{CC}$) yield unbiased estimates of relative risk and cumulative risk parameters, unlike the methods that assume conditional independence of phenotypes given genotypes ($\theta = 1$); these latter methods lead to overestimates of cumulative risk (Table 1). Moreover, we found that the proposed methods yield nearly unbiased estimates of relative risk and cumulative risk parameters even under modest misspecification of the copula model of association. We recommend the use of our method with a Clayton copula model, unless other information indicates a need for a different copula model.

Recently several researchers have raised and studied the issue of ascertainment bias in kin–cohort estimation of cumulative risk from the family histories of case-only probands. In the reported simulations with case-only probands (Table 3), we observe that if the true cumulative risk conferred by a gene is 50%, then the bias from ignoring residual familial correlation ($\theta = 1$) is modest in our analysis based on L_{KC}, but somewhat more important if $L_{KC} \times L_{CC}$ is used. The upward bias can be quite important if the true cumulative risk is 25% (Table 3). Gong and Whittemore (2003) reported little absolute bias in penetrance estimation using the relatives of cases, probably because they assumed that the hazard in noncarriers was known, and also because they studied the situation in which the true penetrance of the gene was 70%.

In principle, one can correct for ascertainment bias even with case-only probands if the correct model for residual familial correlation is known (Table 3). Analyses based on $L_{KC} \times L_{CC}$ yield more precise estimates of relative risks and cumulative risks than analyses based on L_{KC} alone. The case-only proband design, however, has inherent limitations. First, external information on allele frequencies must be used because case-only probands provide very little information on allele frequency. The analysis based on L_{KC} only is more robust to misspecification of the allele frequency parameter than that based on $L_{KC} \times L_{CC}$. A more important limitation of the case-only design is that inferences are not robust to misspecification of the copula model (Table 5), unlike the situation when both case and control probands are available. Thus, unless one has information on residual correlation structure, another approach is needed.

One possibility with case-only probands is to limit estimation to relative risks. For example, the Clayton model with $\theta = 1$ (Table 3), which corresponds to assuming conditional independence given individual genotypes, yielded nearly unbiased relative risk estimates for a low-penetrant common variant and only modest bias for a high-penetrant rare variant. If one has external information on allele frequency and on the age-specific risk of the disease in the general population, $\lambda^*(i)$, as can be obtained from cancer registry data, one can use the age-specific relative risk estimates to obtain age- and genotype-specific hazard rates. For example, for the dominant model, the attributable risk in age interval i is $AR(i) = 1 - [\{f^2 + 2f(1-f)\}\beta_i + (1-f)^2]^{-1}$. The baseline hazard is estimated from $\lambda_0(i) = \{1 - AR(i)\}\lambda^*(i)$, and the hazard for carriers of a dominant mutation is estimated from $\exp(\beta_i)\lambda_0(i)$ by substituting appropriate parameter estimates, $\hat{\beta}_i$ and $1 - A\hat{R}(i)$. Thus, it may be possible to supplement information on relative risk from the case-only proband study to obtain reliable cumulative risk estimates. The same strategy can also be adopted for case–control studies and may yield more precise estimates of cumulative risk parameters than internal estimates of the baseline hazard function. For many diseases, however, reliable external data may not be available, particularly for special subgroups, such as the Ashkenazi Jewish population analyzed in Section 5.

References

Andersen, J. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B* **32,** 283–301.

Begg, C. B. (2002). On the use of familial aggregation in population-based case probands for calculating penetrance. *Journal of the National Cancer Institute* **16,** 1221–1226.

Chatterjee, N. and Wacholder, S. (2001). A marginal likelihood approach for estimating penetrance from kin-cohort designs. *Biometrics* **57,** 245–252.

Chatterjee, N., Shih, J., Hartge, P., Brody, L., Tucker, M., and Wacholder, S. (2001). Association and aggregation analysis using kin-cohort designs with applications to genotype and family history data from the Washington Ashkenazi Study. *Genetic Epidemiology* **21,** 123–138.

Claus, E. B., Schildkraut, E. S., Edwin, I. S., Jr., Berry, D., and Parmigiani, G. (1998). Effect of BRCA1 and BRCA2 on the association between breast cancer risk and family history. *Journal of the National Cancer Institute* **90,** 1824–1829.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiologic studies of familial tendencies in chronic disease incidence. *Biometrika* **650,** 909–917.

Cornfield, J. (1951). A method of estimating cumulative rates from clinical data: Applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute* **11,** 1269–1275.

Efron, B. and Tibshirani, R. J. (1998). *An Introduction to the Bootstrap.* Boca Raton, Florida: CRC Press.

Frank, M. J. (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Mathematicae* **19,** 194–226.

Gail, M. and Chatterjee, N. (2004). Some biases that may affect kin-cohort studies for identified disease genes. In *Proceedings of the Second Seattle Symposium in Biostatistics*, D. Y. Lin and P. J. Heagerty (eds), 175–187. New York: Springer.

Gail, M., Pee, D., Benechou, J., and Carroll, R. (1999a). Designing studies to estimate the penetrance of an identified autosomal mutation: Cohort, case–control, and genotyped-proband design. *Genetic Epidemiology* **16,** 15–39.

Gail, M., Pee, D., and Carroll, R. (1999b). Kin-cohort designs for gene characterization. *Journal of the National Cancer Institute, Monograph* **26,** 55–60.

Genest, C. and MacKay, R. J. (1986). The joy of copulas: Bivariate distributions with given marginals. *American Statistician* **40,** 280–283.

Genest, C. and Rivest, L. P. (1993). Statistical inference procedure for bivariate Archimedean copulas. *Journal of the American Statistical Association* **88,** 1034–1043.

Godambe, V. P. (1991). *Estimating Functions*. Oxford: Oxford University Press.

Gong, G. and Whittemore, A. (2003). Optimal designs for estimating penetrance of rare mutations of a disease susceptibility gene. *Genetic Epidemiology* **24,** 173–180.

Hsu, L., Prentice, R. L., Zhao, L. P., and Fan, J. J. (1999). On dependence estimation using correlated failure time data from case-control family studies. *Biometrika* **86,** 743–753.

Khoury, M. J. and Flander, W. D. (1995). Bias in using family history as a risk factor in case-control studies of disease. *Epidemiology* **6,** 511–519.

Li, H., Yang, P., and Schwartz, A. G. (1998). Analysis of age of onset data from case–control family studies. *Biometrics* **54,** 1030–1339.

Liang, K. Y., Zeger, S. L., and Quaqish, B. (1992). Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society, Series B* **54,** 3–40.

Moore, D. F., Chatterjee, N., Pee, D., and Gail, M. H. (2001). Pseudo-likelihood estimates of the cumulative risk of an autosomal dominant disease from a kin-cohort study. *Genetic Epidemiology* **20,** 210–227.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84,** 487–493.

Patil, G. P. and Rao, C. R. (1978). Weighted distribution and size biased sampling with application to wildlife populations and human families. *Biometrics* **34,** 179–189.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66,** 403–411.

Saunders, C. L. and Begg, C. B. (2003). Kin-cohort evaluation of relative risks of genetic variants. *Genetic Epidemiology* **24,** 220–229.

Shih, J. H. (1998). Modeling multivariate discrete failure time data. *Biometrics* **54,** 1115–1128.

Shih, J. H. and Chatterjee, N. (2002). Analysis of survival data from case-control family studies. *Biometrics* **58,** 502–509.

Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* **51,** 1384–1399.

Struewing, J. P., Hartge, P., Wacholder, S., Baker, S. M., Berlin, M., McAdams, M., Timmerman, M. M., Lawrence, B. C., and Tucker, M. A. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *New England Journal of Medicine* **336,** 1401–1408.

Wacholder, S., Hartge, P., Struewing, J. P., Pee, D., McAdams, M., Lawrence, B. C., and Tucker, M. A. (1998). The kin-cohort study for estimating penetrance. *American Journal of Epidemiology* **148,** 623–630.

Whittemore, A. S. (1995). Logistic regression of family data from case-control studies. *Biometrika* **82,** 57–67.

Whittemore, A. S. and Gong, G. (2003). Re: On the use of familial aggregation in population-based case probands for calculating peretrance. *Journal of the National Cancer Institute* **95,** 76–77.

Zhao, L. P., Hsu, L., Holte, S., Chen, Y., Quiaoit, F., and Prentice, R. L. (1998). Combined association and aggregation analysis of data from case-control family studies. *Biometrika* **85,** 299–315.

Received July 2004. *Revised April* 2005.
Accepted July 2005.